Multimedia Tools and Applications (2022) 81:41899–41910
https://doi.org/10.1007/s11042-021-11473-z

1214: MULTIMEDIA MEDICAL DATA-DRIVEN DECISION MAKING

Mitigating adversarial evasion attacks by deep active learning for medical image classification

Usman Ahmed[1] · **Jerry Chun-Wei Lin**[1] · **Gautam Srivastava**[2,3]

Received: 19 December 2020 / Revised: 27 July 2021 / Accepted: 19 August 2021 /
Published online: 12 July 2022

Abstract

In the Internet of Medical Things (IoMT), collaboration among institutes can help complex medical and clinical analysis of disease. Deep neural networks (DNN) require training models on large, diverse patients to achieve expert clinician-level performance. Clinical studies do not contain diverse patient populations for analysis due to limited availability and scale. DNN models trained on limited datasets are thereby constraining their clinical performance upon deployment at a new hospital. Therefore, there is significant value in increasing the availability of diverse training data. This research proposes institutional data collaboration alongside an adversarial evasion method to keep the data secure. The model uses a federated learning approach to share model weights and gradients. The local model first studies the unlabeled samples classifying them as adversarial or normal. The method then uses a centroid-based clustering technique to cluster the sample images. After that, the model predicts the output of the selected images, and active learning methods are implemented to choose the sub-sample of the human annotation task. The expert within the domain takes the input and confidence score and validates the samples for the model's training. The model re-trains on the new samples and sends the updated weights across the network for collaboration purposes. We use the InceptionV3 and VGG16 model under fabricated inputs for simulating Fast Gradient Signed Method (FGSM) attacks. The model was able to evade attacks and achieve a high accuracy rating of 95%.

Keywords Adversarial attack · IoMT · Medical image analysis · Deep learning

✉ Jerry Chun-Wei Lin
jerrylin@ieee.org

Usman Ahmed
usman.ahmed@hvl.no

Gautam Srivastava
srivastavag@brandonu.ca

[1] Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, 5063, Bergen, Norway

[2] Department of Mathematics & Computer Science, Brandon University, Brandon, Canada

[3] Research Centre for Interneural Computing, China Medical University, Taichung, Taiwan

1 Introduction

Deep Learning (DL) models learn by processing layers of hidden representation and multi-level abstraction. The DL models can achieve generalization by using large-scale data and applied in many domains and applications [1, 18, 19]. The outstanding performance of DL in industry and academia is widely recognized. The medical tasks are being automated using DL methods [20, 35], including cancer detection, tumour classification, vessel segmentation, etc. The high accuracy is impressive; however, models often act as black-boxes with unexplainable complicated layers. Other issues with these models are reported to be *"unreliable"* and *"defenseless"* when exposed to unexpected hand-designed input samples [25]. This mentioned issue makes the models insecure because black-box prediction results in detrimental performance. The DL models are required to be more efficient and robust to any intrusion. With recent advancement, it is still unclear what information must be given as input to the DL methods as results DL method give surety about the robust and secure prediction for the real-world problems [12]. Therefore, explainable learning methods are the latest research trend for the researcher.

Most researchers use medical images and applied their own learning-based methods [36]. Medical images use learned images of vector space for the analysis of specific tasks [14]. However, these methods mislead when applied to adversarial examples, i.e., some training data are used to lack the detection at the testing phase. Many issues in DL theory are caused because of this. It takes the view that the training dataset used in a learning phase should be correlated with the problem domain, and purposeful modification of the data does not occur [37]. One of the common problems in medical imaging occurs is bias in the training data. The training sample bias occurs when samples do not accurately reflect the context in which the learning method will be operated. In the medical domain, an expert to define a problem is used as a data labeling task. The method takes advantage of the *"human in the loop"* to ensure that the experts *(who work with data)* judgment are used with the learning method [32]. It is very much essential to collect samples that are very diverse and close to real-world scenarios [32]. Training set distribution must match the distribution of actual data.

While it is clear that scenarios can offer massive benefits, connected devices sharing information can pose new issues on the Internet of Medical Things (IoMT). There have been lots of improvements in the cyber-infrastructure of IoMT, which can enable activities like the sharing of diagnostic images over large networks and large distances remotely. Medical image processing is at a new dawn in society that needs to be thoroughly researched, developed, and understood. Currently in 4G, moving into 5G, and looking further ahead to 6G, the integration of AI and IoMT will assist in achieving breakthroughs in terms of our ability to analyze and understand the intricate world of medical imaging comprehensively.

1.1 Motivation

With the advent of pandemics and research collaboration across the globe, governments and the healthcare industry invested in the health care economy. The significant investment inevitably creates potential attackers, which might engage in criminal acts to seek profit from manipulating the health care system. For example, with the usage of the distributed data in the healthcare system, an attacker might exploit their examination report to do insurance fraud or claim the false medical reimbursement [24]. On the other hand, the attacker can also use a series of images to cause misdiagnosis of disease and overload the health care system with fake results. This could cause a severe impact on the decision-maker about the patient of certain Geo-locations. The black box decision making of DNN [23], the

misdiagnosis with high confidence is hardly recognized and explainable. The secure, trustable and explainable are crucial [15, 24], as the DNN system increasing adopted in the medical diagnostics, decision support, and pharmaceutical approvals [26]. There are several methods used to fool learning models. The tampered input features fall under the umbrella of deep adversarial learning. Attacks of learning methods are classified into two major categories, i.e. evasion attacks and poisoning attacks. The evasion attacks are performed when the attacker changes the input samples, leading to a miss classification of the legitimate samples. The poisoning attacks are performed during the training phase when an attacker manipulated the training data with handcrafted samples to compromise the learning process.

In this research, we work to deal with adversarial evasion attacks. We used the federated learning approach to share the model across medical institutes. Then the proposed active learning-based system is to defend against adversarial evasion and poisoning attacks. The proposed centroid-based clustering approach constructs the unsupervised feature representation into two classes, i.e., adversarial examples and quality samples (training set distribution). Then, we use the active learning method to train the learning model on the limited samples of the institute. We use the human in the loop to select the sampling for each round of the active learning method. The major motivation of using the input clustering into adversarial samples was to enable each representation to contribute to the training (evading poisoning attacks) of the learning method. Also, the active learning method with the human-in-the-loop methods helps to make the job difficult for any attacker. Hence, creating a robust and secure framework for medical image tasks, in particular, we made the following contributions:

- For medical imaging data, we proposed the embedding model for distributed institute at the edge.
- We proposed a dynamic clustering model that clusters the low dimensional vectors into meaningful regions.
- We investigate the usage of federated learning and dynamic cluster for the medical data. The proposed model improves the performance with deep active learning without transmitting any raw data.

2 Related work

Deep neural networks (DNN) achieve humans like performance over several tasks such as image classification, object detection [36], and image retrieval [4]. With the recent success in CIFAR-10 and ImageNet, the DNN models gain popularity, success, and domain adoption, including in the health care industry. DNN become a powerful tool for analyzing and diagnosing disease for medical image processing tasks, such as diabetic retinopathy detection,[1] cancer diagnosis [13], and organ segmentation [28]. However, the high performance of these larger and deeper networks has issues regarding the user's trustworthiness and safety. In a recent study, it is found that models are highly vulnerable to carefully crafted adversarial examples (for attacks), i.e., the slight change in the input data or instance distribution, the result of DNN can be changed. In addition to the wrong output, the prediction confidence is also on the higher side [16, 33]. This means that model is classifying the instance into wrong output with higher confidence which raised issues for the safety of the user in critical applications such as medical diagnosis [15], autonomous driving [14], and

[1] https://www.kaggle.com/c/diabetic-retinopathy-detection

action analysis [9]. The existing methods focus on the adversarial machine learning research on natural image analysis tasks, and the health care image domain is still an open research area. The characteristic of the medical images is different from the normal image, for example, biological texture and shape. Ravi et al. [27] presented a novel approach to detect randomly generated domain names and domain name system (DNS) homograph attacks without the need for any reverse engineering or using nonexistent domain (NXDomain) inspection by DL models. Maarouf et al. [21] investigated the effectiveness of different evasion attacks and see how resilient machine and deep learning algorithms. Amich and Eshete [3] introduced a novel framework that harnesses explainable ML methods to guide high-fidelity assessment of ML evasion attacks. A recent study confirmed that adversarial attacks on the medical deep learning system could cause the problem and make them vulnerable [12, 15, 24, 25].

Recent work has analyzed the robustness of the deep models on medical images and focuses on the testing of the framework [15]. Another work focus on testing the vulnerability of medical deep learning models [24]. The testing showed that the classification accuracy drops from 86% to 0% on adversarial examples. The robustness and vulnerability of the model are tested against the adversarial standards on classification or segmentation tasks [24]. Also, the authors [24] conducted the experiment by small perturbations and observed a small change in the input variable, and the performance drops across different models. However, still, it should be evaluated how medical image attacks can also be crafted as raw image attacks? If not, then why this happened? In this study, we explain the detection and prevention of the attacks both at training, validation, and testing experiments.

Several DNN methods are used for diagnostic and prognostic biomarkers [30]. Mainly, DNN required diverse and extensive training sets, which is challenging to acquire in the medical domain. Collaboration among institutes faces data privacy and ownership issues. A federated learning framework helps to provide data-private multi-institute collaboration. The method [30] investigate the data distribution as assessing the model quality and learning patterns. Clinical adoption of federated learning helps to trained models and achieved high precision in personalized diagnostics. The model uses the federated learning method for data privacy. However, the adversarial method and detection are not evaluated.

The study, [38], addressed the over-fitting issues in the deep learning model. The analysis found that the data distribution is an essential factor as model training on one institute was performing purely on the testing sample of another institute. The reason is that the sampling and distribution of testing were different from the training instances. It is noted that DNN in the medical imaging domain is specifically associated with institutional biases. These models can achieve high accuracy but are not able to generalize the specific diagnosis to external institutions or even intra-department healthcare centres [38]. The data distribution should be diverse enough to be effective. In this research, we handle this problem by using active learning methods and federated learning methods. This is a natural way of collaboration, and the technique helps increase data diversity among multiple institutions.

Mostly, data is stored in a centralized location for collaboration. The number of data sharing repositories still exists for medical fields, e.g., radiology, pathology, and genomics [2, 6, 10]. Collaborative data sharing (CDS) have several issues, i.e., patient privacy, data ownership, and international laws [8]. Consequently, patterns extracted from the diverse populations are required to be removed from multiple sources and institutions. The number of collaborative learning approaches enables training models among different institutions without sharing patient data [7].

Federated learning (FL) is a collaborative learning approach without sharing the data [22]. The FL train a machine learning model, then transfer the model across the network

to a server to be aggregated into the consensus model. After aggregating the model on the server, it is shared with all collaborated sources for further training and usage. Each iteration process, training of the model in parallel, apply aggregation and distribute the model weights across the network called federated rounds [22]. Fl was introduced by Google [22] as federated averaging and later applied Google keyboard for auto text completion task [5]. Chang *et al.* [7] proposed a collaborative learning method for the medical model, where institute training is done in parallel, and then model weights transferring was performed incrementally and in cyclic order. In incremental learning, each institute is trained, and then model weights are transfer to the next one [7]. In cyclic learning, a fixed number of training runs was performed for each institute and repeated through the number of the institute [29]. The cyclic order and limited epochs per institute help the cyclic learning model enable gradual progress and achieve better results than other models [29]. However, adversarial attacks are required to be avoided, and the mechanism is required to handle the situation.

Algorithm 1 Adversarial detection based federated averaging method.

INPUT: T images data, R are the number of rounds, n_K are the local training epochs to minimize loss $\mathcal{L}_k\left(X_k; \phi^{(t-1)}\right)$ for client K

OUTPUT: Optimize weights

$Weights \leftarrow \phi^{random}$ ▷Initialize weights randomly
for *all* $r \in R$ **do**
  for *all client* $\in K$ **do**
    $Receive\left(\Delta\phi_k^{(r)}, n_k\right)$
    for *all* $i \in N$ **do**
      if $i \leq N$ **then**
        $k \leftarrow k + 1$
        $SV_k \leftarrow SIM(T_i)$
      else
        $k \leftarrow 1$
        $SV_k \leftarrow SIM(T_i)$
      end if
      repeat until convergence
      Find the mean centroid for each cluster
      for *all* $k \in K$ **do**
        $similarity \leftarrow SIM(M_k, SM_N)$
        Re-assign(S, M)
        Reduce (S, M)
      end for
      $Uncertainty_{pool} \leftarrow Entropy\ ()$
      $Human_{Annotation}()$
      $Updatetrainingset()$
      $Retrain(R, epochs = 10)$
      $Send\ \phi^{r-1}$
    end for
  end for
  $\phi_k^{(r)} \leftarrow \phi^{(r-1)} + \Delta\phi_k^{(r)}$
  $\phi^{(r)} \leftarrow \frac{1}{\sum_k n_k} \sum_k (n_k \cdot \phi_k^{(r)})$ ▷ Aggregate( )
end for
Return $\phi^{(r)}$

3 Centroid active learning method (CALM)

To analyze the adversarial samples $\boldsymbol{x}$,$'$ and used by the feature extractor f_θ, we designed the unsupervised active learning method g_ϕ based on centroid clustering method. The method goal is to exclude the unseen adversarial samples $\boldsymbol{x}' \in \mathcal{X}'$, and extract to prevent $\boldsymbol{x} \in \mathcal{X}$ from being removed from the training distribution. The proposed clustering method is unsupervised that does not require adversarial density or type of adversarial attacks. Additionally, the model prediction is first selected based on the uncertainty-based pool selection. After that, the select class, along with the confidence, was validated by the human annotator.

A federated learning method is mentioned the Fig. 1, where customized weights are moved globally (though all the institutes are connected over a network). After receiving the optimized weights, the institute trained the model locally. In our proposed approach mentioned in Fig. 2, unlabeled samples first clustered into two groups. Given an input image, we are required to detect if the example is normally distributed or some permutation is added to it. Initially, we use the pre-trained network to extract the feature vector. Then initialize with random distribution (initialization), each with the mean centroid. We reassign the cluster by computing the similarity determined by centroid based on the feature extracted embedding *(Algorithm 1 — lines 6 to 12)*. Then, the next iteration repeats until the centroid converges is not changed. The clusters are just collection; we can define the centroid of the cluster by computing the average of the image belonging to each cluster; thus, if $S_1, S_2, \ldots S_N$ are images belonging to cluster and centroid *(Algorithm 1 — lines 13 to 19)*. All the unlabeled images in the cluster except for the cluster centroid images are being compared. Thus, we have sample images $S_1, S_2, \ldots S_N$ and we want similarity

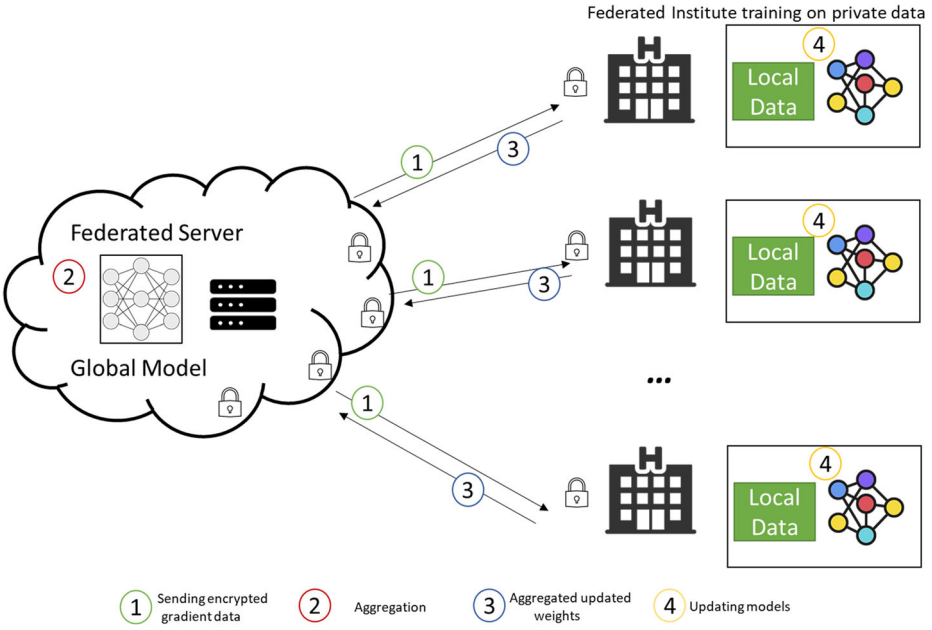

Fig. 1 Federated learning in medical imaging. The server communicates with the multiple institutes for the weights and gradient from the local model. The local models are aggregated and shared to achieve high performance

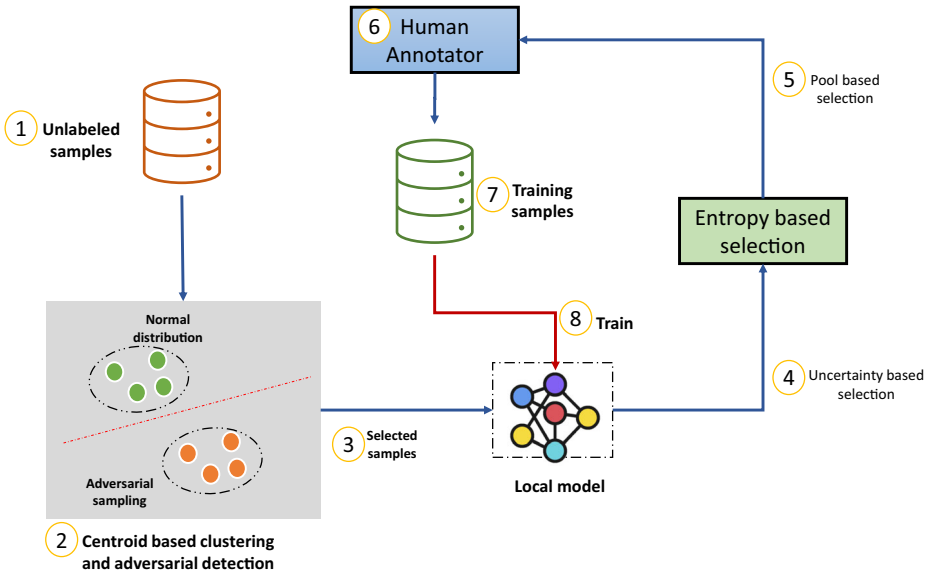

Fig. 2 A cluster-based adversarial attacks detection and centroid based model evasion

between clusters and images appearing in the cluster, we determine the cluster centroid using the $S_1 \cup S_2 \cup \ldots \cup S_{G-1} \cup_{G+1} \cup \ldots \cup S_N$. SM_N is the embedding belonging to cluster K $N = 1 \ldots j$ where j is the number of images belonging to cluster K. For similarity among vector, we used the cosine similarity mentioned as in the (1) and *(Algorithm 1 — line 16)*. Given two vectors t, e, where t represents image embedding and t is the centroid. We compute cosine similarity to compare centroid and images embedding two vectors.

$$\sim (\mathbf{t}, \mathbf{e}) = \frac{\text{te}}{\|\mathbf{t}\| \|\mathbf{e}\|} = \frac{\sum_{i=1}^{n} \mathbf{t}_i \mathbf{e}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{t}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{e}_i)^2}} \tag{1}$$

After the adversarial detection, we used the trained model to predict the class of the raw images. The model takes advantage of the optimized weights and gradient to learn from a limited set of examples. The initial training set contains a small amount of data and an entropy-based model to decide which points to select for inclusion in the training set. The entropy-based model in this paper selects the number of a point depending upon the pool size *(Algorithm 1 — lines 20 to 22)*. The pool point is the separate set that is updated each cycle. It includes the selected point in the training set, and we train an alternative model on the new points *(Algorithm 1 — lines 22 to 23)*. The repetition of the steps helps to increase with training-set and meaningful points over time. The human annotator then validates the select data points by comparing the confidence score and the output class. The developed method can thus help to reduce data annotation tasks and generalize the machine learning system.

The institute (connected with the server) can benefit from local data without sharing it across the network mention in Fig. 1. We use this concept in this research; the server is an online computational efficient system where the institute is local hospitals in this way system able to train intelligent system by keeping the privacy of any individual patient. We mention the framework for the multiple institutes in Fig. 1. Each client data exhibits

non-overlapping distribution, local model (model at institute end), and database. The data distribution is varied among the client randomly, and adversarial examples are also added. We mention that only the deep learning model is shared among the node; the clustering is performed locally by each institute. Local institutes owners do adversarial attack detection based on the local distribution. However, the feature extractor helps to create a learned embedding clustering purpose. This helps the dynamic cluster method to form the separable margin of normal images and adversarial examples. In the proposed model, the institute's data is stored in local database. On each iteration, the model is trained locally, and then weights/gradient are transferred to the server *(Algorithm 1 — lines 3 to 9)*. The server receives the model and then aggregates the weights of all the clients *(Algorithm 1 — lines 5 to 6)*. This helps the model learn the diverse institute data and transfer it to all the client connections. The purpose of the transfer is that it helps the model to improve the predictive ability and move towards the generalized model. The global model then uses the aggregate weights to update the global model *(Algorithm 1 — lines 8 to 9)*. We used *Federeated Averaging* method for the model aggregation [22]. The server performs the aggregation, and then the updated model is shared with the institute with global weights. After a particular iteration of the rounds, clients reach the convergence point. Then, the final global sharing is performed among the clients. During the experimental evaluation, we set the round to 10 and epochs per institute to 20. The institute can select the global aggregated model or the best local iteration-based data. After getting the gradient and weights, we perform the dynamic cluster. For optimization and hyper-parameter tuning, we set the initial stopping patience value to 10.

For our experimental analysis, we used the retinal OCT image dataset [17]. In total, it contains 84,495 images from 4,686 patients with four output classes (normal, drusen, diabetic macular edema (DME), and choroidal neovascularization (CNV). We used only 1,000 images of data (800 training and 200 testing purpose). To balance the output class, we used the under-sampling method. We randomly distributed the samples across four institutes. Additionally, we added adversarial samples to test the robustness of the model. The FGSM method [16] is used to generate a sample with permutation interval set to [0.002, 0.004]. For each institute data, we create 1:1 adversarial samples. The VGG-16 [31] and inceptionV3 [34] model was used to train the network and ImageNet [11] was used as a feature extractor. The model was performed for ten epochs and 20 rounds each. We evaluated the model

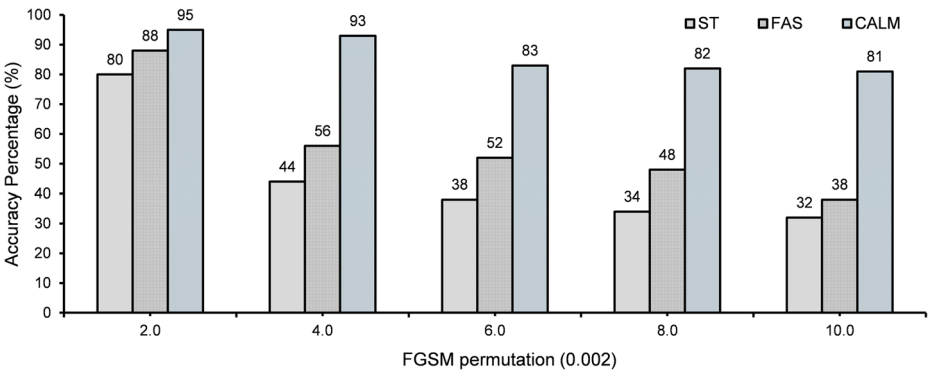

Fig. 3 The active learning method comparison with standard training method of federated averaging (ST), federated averaging with adversarial samples (FAS) [5], federated averaging by using the proposed framework evasion method, i.e., Centroid active learning method (CALM)

Table 1 True classify samples by using the Centroid active learning method (CALM) results on with adversarial samples

Normal	**39**	0	0	0
CNV	0	**41**	0	1
DME	5	2	**31**	1
DRUSEN	1	3	0	**36**
	Normal	CNV	DME	DRUSEN

with the standard training method of federated averaging (ST), federated averaging with adversarial samples (FAS), federated averaging by using the proposed framework evasion method, i.e., Centroid active learning method (CALM).

3.1 Validation of adversarial attacks

An attacker can deliberately fabricate the input application to force the classification algorithm to produce the desired output (evasion attack). To validate the resilience of the proposed centroid active learning method against the adversarial evasion attacks, we experimented using the FGSM model employing the fabricated inputs. These fabricated inputs (to mimic an adversarial evasion attack) are produced by making slight changes in feature vectors of known images. We evaluated the performance of the proposed model to mitigate evasion attacks by making permutation values under different thresholds. Most of the existing deep learning detection techniques are vulnerable to adversarial evasion attacks because the bit change in the input feature vector results in evasion from the underlying classification model. To evaluate the proposed model against such fabrication in the input feature vector, we used to cluster the method and select quality features to train the network purely. This results in evasion of model poisoning as well as model evasion. We tested our proposed method using the fabricated feature vectors by employing different DL architectures.

It is evident from Fig. 3, that the proposed model achieved 95% to 81% accuracy respectively for all fabricated data. All other models achieved lower accuracy with a maximum of 88 and a minimum of 32%. Although these techniques achieve high accuracy for the limited number of permutations, authors have not provided any countermeasure to mitigate such attacks. In comparison to these techniques, our proposed method achieved a remarkable 85% against adversarial examples. The accuracy of different ensembles for 10-bit fabricated data can be seen in Fig. 3, which displays the same trend regarding the performance of different ensembles. The results obtained from these experiments validate the fact that the proposed adversarial detection method is resilient to adversarial evasion attacks. The proposed approach can detect an attack on medical imaging and the fabricated samples with high accuracy.

Tables 1 and 2 show the performance of InceptionV3 on adversarial data. The results indicate that the performance of the proposed model is significantly high. Both methods performed well for the provided distinct data. However, when the robustness of the model is detected under fabricated samples, then VGG16 tends to perform very low.

Table 2 CALM accuracy comparision with VGG16 and InceptionV3 model

Accuracy	VGG16	InceptionV3
With adversial examples	0.88	0.92
Without adversial examples	0.94	0.96

Table 3 Model accuracy comparison with local and federated learning method

	Test			
Client	1	2	3	4
(a)- Local				
1	**0.78**	0.62	0.02	0.02
2	0.06	**0.91**	0.01	0.02
3	0.01	0.03	**0.92**	0.02
4	0.012	0.001	0.01	**0.88**
(b)- Federated				
1	**0.82**	0.62	0.02	0.02
2	0.06	**0.97**	0.01	0.02
3	0.01	0.03	**0.96**	0.02
4	0.012	0.001	0.01	**0.92**

Table 3 shows the accuracy comparison of the Inceptionv3 model by using the local training data alone and a federated learning method. The performance of the locally best model (having high validation score) is improved after the usage of shared weights of federated learning. We also observed a remarkable improvement diagonal mean of the client's test accuracy. It is concluded that model trends to be performed locally and globally under the adversarial samples.

4 Future work

In future work, we will employ new training strategies to select the subset of client training data to improve the frequency of iterations, mini-batch sampling, and hyper-tuning models. We also want to employ deep adversarial or generative models to evade the system. Differential privacy methods can also be integrated to improve the effectiveness and increase the quality of training to minimize bias inactivity to the data being sampled. For the possible weight aggregation method, the evolutionary computation method can also be adopted that uses the institute local model weights to further optimize the global model aggregation.

5 Conclusion

In conventional healthcare systems and now moving into our connected world of IoMT, collaboration among institutes is considered to be highly effective to enable distributed healthcare services. Clinical trials dataset has a built-in problem that includes a lower number of training instances, sampling biases, presences of outliers, and class imbalance. The deep learning-based model trained on the limited data results in overfitting and under-formed when a process on new unseen data. In this research, we proposed a data collaboration method using the federated learning approach for the healthcare industry. The model first detects the adversarial examples and then predicts the outcome of the select images after the entropy base uncertainty model can reduce the set of examples for the human annotation task. In the end, the human annotator takes the input images, output

class, and confidence score to validate the sample. In this way, the adversarial attack can be evaded to poison the model and misclassify the input images. The obtained results support our decisions regarding deep neural network training to detect and mitigate evasion attacks. Moreover, it validates itself as a resilient model against adversarial evasion attacks by achieving 95% accuracy against fabricated inputs.

Funding Open access funding provided by Western Norway University Of Applied Sciences.

References

1. Ahmed U, Lin JCW, Srivastava G (2021) Privacy-preserving deep reinforcement learning in vehicle adhoc networks. IEEE Consum Electron Mag
2. Aldape K et al (2018) Glioma through the looking GLASS: molecular evolution of diffuse gliomas and the glioma longitudinal analysis consortium. Neuro-Oncol 20(7):873–884
3. Amich A, Eshete B (2021) Explanation-guided diagnosis of machine learning evasion attacks. arXiv:2106.15820
4. Bai X, Yan C, Yang H, Bai L, Zhou J, Hancock ER (2018) Adaptive hash retrieval with kernel based similarity. Pattern Recogn 75:136–148
5. Bonawitz K, Eichner H, Grieskamp W, Huba D, Ingerman A, Ivanov V, Kiddon C, Konečnỳ J, Mazzocchi S, McMahan HB et al (2019) Towards federated learning at scale: System design. arXiv:1902.01046
6. Borovec J et al (2020) ANHIR: Automatic Non-rigid histological image registration challenge. IEEE Trans Med Imaging 39(10):3042–3052
7. Chang K, Balachandar N, Lam C, Yi D, Brown J, Beers A, Rosen B, Rubin DL, Kalpathy-Cramer J (2018) Distributed deep learning networks among institutions for medical imaging. J Am Med Inform Assoc 25(8):945–954
8. Chen M, Qian Y, Chen J, Hwang K, Mao S, Hu L (2020) Privacy protection and intrusion avoidance for cloudlet-based medical data sharing. IEEE Trans Cloud Comput 8(4):1274–1283
9. Cheng Y, Lu F, Zhang X (2018) Appearance-based gaze estimation via evaluation-guided asymmetric regression. In: Computer vision. Springer, pp 105–121
10. Davatzikos C et al (2020) AI-Based prognostic imaging biomarkers for precision neuro-oncology: the reSPOND consortium. Neuro-Oncol 22(6):886–888
11. Deng J, Dong W, Socher R, Li L, Li K, Li FF (2009) Imagenet: a large-scale hierarchical image database. In: IEEE Conference on computer vision and pattern recognition, pp 248–255
12. Ding X, Zhang S, Song M, Ding X, Li F (2021) Toward invisible adversarial examples against DNN-based privacy leakage for internet of things. Internet Things J 8(2):802–812
13. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542(7639):115–118
14. Eykholt K, Evtimov I, Fernandes E, Li B, Rahmati A, Xiao C, Prakash A, Kohno T, Song D (2018) Robust physical-world attacks on deep learning visual classification. In: Conference on computer vision and pattern recognition
15. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS (2019) Adversarial attacks on medical machine learning. Science 363(6433):1287–1289
16. Goodfellow IJ, Shlens J, Szegedy C (2015) Explaining and harnessing adversarial examples. In: International conference on learning representations
17. Kermany DS et al (2018) Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 172(5):1122–1131.e9

18. Lin JCW, Shao Y, Zhou Y, Pirouz M, Chen HC (2019) A bi-lstm mention hypergraph model with encoding schema for mention extraction. Eng Appl Artif Intell 85:175–181
19. Lin JCW, Shao Y, Djenouri Y, Yun U (2021) Asrnn: a recurrent neural network with an attention model for sequence labeling. Knowl-Based Syst 212:106548
20. Lyu Z, Wang Z, Luo F, Shuai J, Huang Y (2021) Protein secondary structure prediction with a reductive deep learning method. Front Bioeng Biotechno 9:687426
21. Maarouf R, Sattar D, Matrawy A (2021) Evaluating resilience of encrypted traffic classification against adversarial evasion attacks. arXiv:2105.14564
22. McMahan B, Moore E, Ramage D, Hampson S, Arcas BA (2017) Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics, pp 1273–1282
23. Niu Y, Gu L, Lu F, Lv F, Wang Z, Sato I, Zhang Z, Xiao Y, Dai X, Cheng T (2019) Pathological evidence exploration in deep retinal image diagnosis. AAAI Conf Artif Intell 33:1093–1101
24. Paschali M, Conjeti S, Navarro F, Navab N (2018) Generalizability vs. robustness: Investigating medical imaging networks using adversarial examples. In: Medical image computing and computer assisted intervention, pp 493–501
25. Paranjape JN, Dubey RK, Gopalan VV (2020) Exploring the role of input and output layers of a deep neural network in adversarial defense. In: International conference on computing and data science, pp 114–118
26. Pien HH, Fischman AJ, Thrall JH, Sorensen A (2005) Using imaging biomarkers to accelerate drug development and clinical trials. Drug Discov Today 10(4):259–266
27. Ravi V, Alazab M, Srinivasan S, Arunachalam A, Soman KP (2021) Adversarial defense: DGA-based botnets and DNS homographs detection through integrated deep learning. IEEE Trans Eng Manag
28. Roth HR, Lu O (2015) Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 556–564
29. Roth HR, Chang K, Singh P, Neumark N, Li W, Gupta V, Gupta S, Qu L, Ihsani A, Bizzo BC et al (2020) Federated learning for breast density classification: a real-world implementation. In: Domain adaptation and representation transfer, and distributed and collaborative learning, pp 181–191
30. Sheller MJ et al (2020) Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Scient Rep 10(1):12598
31. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
32. Stapor K, Ksieniewicz P, García S, Woźniak M (2021) How to design the fair experimental classifier evaluation. Appl Soft Comput 104:107–219
33. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2013) Intriguing properties of neural networks. arXiv:1312.6199
34. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: IEEE conference on computer vision and pattern recognition, pp 1–9
35. Wang Z, Cai B (2021) COVID-19 Cases prediction in multiple areas via shapelet learning. Appl Intell 1–12
36. Wang C, Bai X, Wang S, Zhou J, Ren P (2019) Multiscale visual attention networks for object detection in VHR remote sensing images. IEEE Geosci Remote Sens Lett 16(2):310–314
37. Yu Z, Zhou Y, Zhang W (2020) How can we deal with adversarial examples? In: International conference on advanced computational intelligence, pp 628–634
38. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK (2018) Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Med 15(11):e1002683